Technical Disclosure Commons

Defensive Publications Series

March 2020

Mixed Language Speech Recognition

Ágoston Weisz

Herbert Jordan

Konrad Jan Miller

Mixed Language Speech Recognition

ABSTRACT

Mixed language speech, e.g., speech in two or more languages, is common in many geographies and application domains. A mixed language voice command to a virtual assistant is likely to be misinterpreted by conventional speech recognition techniques that are based on a single language model. This disclosure describes techniques to combine multiple language models to perform multilingual speech recognition of mixed language speech and commands.

KEYWORDS

- Speech recognition
- Mixed language query
- Spoken query
- Language model
- Phonetic model
- Virtual assistant
- Smart speaker

BACKGROUND

Mixed language speech, e.g., speech in two or more languages, is common in many geographies and application domains. A mixed language voice command to a virtual assistant, e.g., “play 99 luftballons on the online music store,” or “navigate to the Dorfplatz in Feldkirch,” which includes both English and German phrases is often misinterpreted by conventional speech techniques that are based on a single language model.

In the above examples, the frame of the query is in English, while relevant elements are in a different language, in this case, German. A mixed-language query of an inverse type, e.g.,

where the frame-language is German and the elements are in English, e.g., “spiel orange is the new black auf streaming-movie provider” is also possible, and can be provided, e.g., by a German-speaking user accessing English-titled content.

A multilingual person fluent in both languages is likely to use native pronunciations for the constituent languages of their speech. However, speech recognition techniques generally utilize a single language model, leading to misrecognition of the spoken query.

DESCRIPTION

This disclosure describes techniques to combine multiple language models to improve speech recognition of mixed language speech and commands.

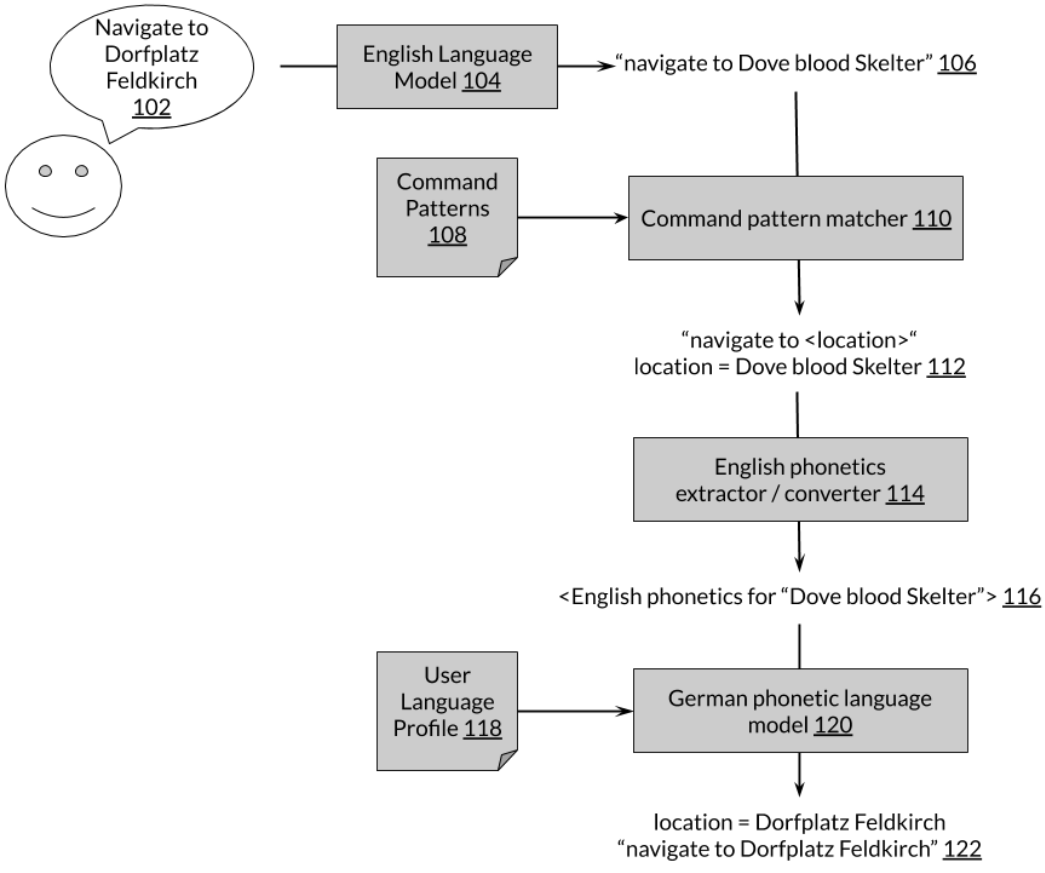

Fig. 1: Mixed-language speech recognition

Per the techniques of this disclosure, a user establishes a user language profile in which the user declares their primary language that is used for the language of the frame of their utterances, and one or more secondary languages that may be used for the elements of their utterances. With user permission, the primary and secondary languages of the user can also be inferred from user history, e.g., from the languages used during searches.

Fig. 1 illustrates mixed-language speech recognition, per the techniques of this disclosure. A user that has set their primary language as English and secondary language as German utters an English-German command, "Navigate to Dorfplatz Feldkirch (102)," which is received by a virtual assistant. An English language ASR model (104) is used to transcribe the user's utterance (106). The English parts of the utterance are recognized correctly; the German parts are incorrect.

The English transcription "dove blood skelter" is determined as implausible, and hence incorrect, based, e.g., on language model scores, audio model scores, speech hypothesis scores, etc. The issued voice command is compared with a set of supported command patterns (108), where sections of the pattern serve as placeholders. In this example, the command matches the pattern "navigate to <*location*>" (112), e.g., the placeholder is <*location*>, presently transcribed incorrectly as "dove blood skelter." The command pattern matcher (110) can be, for example, a trained machine-learning model that can predict the components of a query.

Phonetics of the possibly misrecognized entity are looked up using an English-language phonetics extractor-converter (114). The extracted English phonetics (116) are looked up in the user's secondary language (German, in this case) phonetic model (120), specified in the user language profile (118). The secondary language phonetic model generates a replacement for the placeholder phrase, e.g., location is identified as "Dorfplatz Feldkirch (122)." A final pass of

speech recognition is executed to rescore the presently generated hypotheses to select the better of the mixed-language and the single-language transcriptions of the user command.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's spoken language abilities, preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level) so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques to combine multiple language models to perform multilingual speech recognition of mixed-language speech and commands.

REFERENCES

[1] Sung, Yun-Hsuan, Francoise Beaufays, Brian Strope, Hui Lin, and Jui-Ting Huang. "Recognizing speech in multiple languages." U.S. Patent 9,129,591, issued September 8, 2015.